# CTAKESProposal

## cTAKES Proposal

The following is a proposal for a new top-level project within the ASF.

### Abstract

cTAKES: (clinical Text Analysis and Knowledge Extraction System) is an natural language processing tool for information extraction from electronic medical record clinical free-text.

### Proposal

cTAKES comprises a collection of components and tooling written in Java specifically trained for the clinical domain, and creates rich linguistic and semantic annotations that can be utilized by clinical decision support systems & clinical research.

### Background

The development of cTAKES started in 2006 by a team of physicians, computer scientists and software engineers at the Mayo Clinic. The development team was led by Dr. Guergana Savova & Dr. Christopher Chute. cTAKES is released open source under an Apache v2.0 license. This system was deployed at Mayo and is currently an integral part of their clinical data management infrastructure and has processed in excess of 80 million clinical notes. Currently, the core development team is co-located at Mayo Clinic and Children's Hospital Boston following Dr. Savova's move to Children's Hospital Boston in early 2010. Additional collaborations with external groups at University of Colorado, Brandeis University, University of Pittsburgh, University of California at San Diego continue to extend the capabilities of cTAKES into areas such Temporal Reasoning, Clinical Question and Answering, and coreference resolution for the clinical domain. In 2010, cTAKES was adopted by the I2B2 program and is a central component of the SHARP Area 4. The current cTAKES components include:

- Sentence boundary detector
- Rule-based tokenizer to separate punctuations from words
- Normalizer
- Context dependent tokenizer
- Part-of-speech tagger
- Phrasal chunker
- Dictionary lookup annotator and normalization to an ontology
- Context annotator
- Negation detector
- Dependency parser
- Constituency parser
- Semantic Role Labeler
- Coreference resolver
- Module for the identification of patient smoking status
- Drug mention annotator

### Rationale

We believe there is a clear gap between cutting edge technologies developed out of research labs and in the clinical practice. We believe that moving cTAKES development to the Apache development community will lead to faster innovation, better integration with other open source software, and broader adoption of cTAKES within clinical institutions and improve our healthcare system. We believe that having cTAKES on Apache will encourage the development of a basic set of open source components that will jumpstart these developers' efforts.

### Initial Goals

The initial goals of the proposed project are:

- Bring the community together at the ASF and make the development process transparent for them
- Write user documentation about all major components
- Automated build/continuous integration
- Automate regression tests
- Produce an Incubating release

### Current Status

#### Meritocracy

Some of the initial committers are familiar with Apache's idea of meritocracy, others aren't. We will get everybody on the same level as part of the incubation process.

#### Community

cTAKES already has a considerable user base, both in industry and academia.

### Core Developers

See the initial committer list.

### Alignment

cTAKES has tie-ins with several existing Apache projects. We have been building our components using the UIMA framework. We are also reusing existing Apache projects such as Lucene, Solr, Maven. We expect these collaborations to strengthen further after our move to Apache and experiment with other projects under the Lucene umbrella such as Hadoop and Mahout. Another obvious connection exists to some of the projects under the OpenNLP umbrella.

## Known Risks

### Orphaned products

The project has been around for quite a number of years already, it has a well-established user community and a diverse set of committers.

### Inexperience with Open Source

cTAKES has been an open source project for many years. Many of the developers are already familiar with both open source in general and the ASF in particular.

### Homogenous Developers

The current group of developers is very diverse and spans globally and across multiple institutions.

### Reliance on Salaried Developers

Most of the developers are not paid to work specifically on cTAKES, so there is little reliance on salaried developers.

### Relationships with Other Apache Products

NLP is often used in search and other algorithms that work with unstructured data, thus cTAKES is likely to be useful to the Lucene and Solr communities. It also aligns nicely with both Mahout and UIMA as well as OpenNLP.

### A Excessive Fascination with the Apache Brand

We think the project aligns nicely with the goals of the ASF to disseminate source code to the public free of charge. Clinical NLP has long been the subject of cutting edge research, but is often lacking in community and shared knowledge. We believe that by bringing cTAKES to the ASF, the Apache brand will help deliver clinical NLP capabilities to a much larger audience and likewise a cutting edge project like cTAKES can further the ASF brand by providing users with tried and true, as well as new, natural language processing capabilities.

## Documentation

- https://wiki.nci.nih.gov/display/VKC/cTAKES+2.0
- http://en.wikipedia.org/wiki/CTAKES

## Initial Source

The source code is maintained in SVN on SourceForge: cTAKES: http://sourceforge.net/projects/ohnlp/

## Source and Intellectual Property Submission Plan

The cTAKES source code is already open source under the AL 2.0.

## External Dependencies

| Library | License | | | Description |
|---|---|---|---|---|
| libsvm | BSD | | | Machine Learning Library |
| UIMA | AL 2.0 | | | Unstructured Information Management Architecture |
| Lucene Core | AL 2.0 | | | Plain Text Search Engine Library |
| OpenNLP | AL 2.0 | | | General Purpose Natural Language Processing Library |

| | | | | |
|---|---|---|---|---|
| HSQLDB | BSD | | | In Memory DB |
| JDOM | Apache Style | | | Java XML Manipulation Libraryv |
| Open AI FSM | Apache Style | | | Finite State Machines Toolset |

# Cryptography

cTAKES neither provides nor uses any cryptography.

# Required Resources

## Mailing lists

- ctakes-dev
- ctakes-private
- ctakes-user
- ctakes-commits

## Subversion Directory

https://svn.apache.org/repos/asf/incubator/ctakes

## Issue Tracking

Jira: cTAKES

## Other Resources

# Initial Committers

| Name | Email | | | CLA |
|---|---|---|---|---|
| Pei J Chen | pei.chen@childrens.harvard.edu | | | yes |
| Sean Finan | sean.finan@childrens.harvard.edu | | | no |
| Guergana K. Savova | guergana.savova@childrens.harvard.edu | | | no |
| James J Masanz | masanz.james@mayo.edu | | | no |

# Affiliations

# Sponsors

## Champion

Jörn Kottmann

## Nominated Mentors

- Jörn Kottmann
- Grant Ingersoll
- Chris A Mattmann

## Sponsoring Entity

The Apache Incubator